11 Verfahren der lexikometrischen Analyse von Textkorpora

IRIS DZUDZEK, GEORG GLASZE,
ANNIKA MATTISSEK, HENNING SCHIRMEL

Lexikometrische Verfahren untersuchen die quantitativen Beziehungen zwischen lexikalischen Elementen in geschlossenen Textkorpora, d. h. in Textkorpora, deren Definition, Zusammenstellung und Abgrenzung klar definiert ist und die nicht im Laufe der Untersuchung verändert werden. Im Rahmen diskursorientierter Ansätze können diese Verfahren genutzt werden, um Rückschlüsse auf diskursive Strukturen und deren Unterschiede zwischen verschiedenen Kontexten wie bspw. Entwicklungen über die Zeit zu ziehen. Ziel lexikometrischer Verfahren in der Diskursforschung ist es also, großflächige Strukturen der Sinn- und Bedeutungskonstitution in Textkorpora zu erfassen (allgemein zur Lexikometrie und korpusbasierten Verfahren[1] der Linguistik s. Maingueneau 1991: 48ff.; Fiala 1994; Bonnafous und Tournier 1995; Lebart, Salem und Berry 1998; Marchand 1998; Teubert 2005; Baker 2006; Lemnitzer und Zinsmeister 2006; Scherer 2006).[2]

Die Verfahren der Lexikometrie und der Korpuslinguistik wurden innerhalb der Sprachwissenschaften entwickelt. Ihre konzeptionellen Grundlagen liegen in der de Saussure'schen Linguistik und zumindest teilweise auch in der Radikalisierung der de Saussure'schen Ansätze im

1 Während in der französischsprachigen Wissenschaftslandschaft eher von *léxicométrie* oder *statistique textuelle* gesprochen wird, ist in englisch- und deutschsprachigen Publikationen die Rede von *corpus linguistics* bzw. Korpuslinguistik.

2 Informationen zur Korpuslinguistik bieten darüber hinaus die Internetseiten http://www.corpus-linguistics.de (Zugriff: 2.2.2007) sowie http://www.bubenhofer.com/korpuslinguistik (Zugriff: 2.2.2007).

Poststrukturalismus sowie den diskurstheoretischen Überlegungen Foucaults. Dabei ermöglichen lexikometrische Verfahren, gerade auch die *Unterschiedlichkeit* von Verweisstrukturen und damit der Bedeutungen von einzelnen Wörtern und Zeichenverkettungen in unterschiedlichen diskursiven Formationen zu erfassen.

Während bei Verfahren der quantitativen Inhaltsanalyse mit der Kategorisierung und Kodierung von Textabschnitten wichtige Teile der Interpretation i. d. R. an den Anfang der Untersuchung gestellt werden (s. Exkurs), steht bei der Lexikometrie die Herausarbeitung quantitativer Beziehungen zwischen lexikalischen Elementen innerhalb eines gegebenen Textkorpus im Vordergrund – der Schwerpunkt der Interpretation wird im Forschungsprozess damit tendenziell nach hinten an das Ende des Forschungsprozesses verlagert. Dabei handelt es sich jedoch „nur" um eine Verlagerung des Schwerpunktes, da die Formulierung der Fragestellung sowie die Zusammenstellung, Definition und Abgrenzung der geschlossenen Korpora immer bereits interpretative Entscheidungen erfordern. Die eigentliche Interpretation der Ergebnisse erfolgt aber erst, nachdem die Ergebnisse der korpuslinguistischen Analysen vorliegen.

Innerhalb der lexikometrischen Verfahren lassen sich zwei Herangehensweisen unterscheiden: Als *corpus based* werden korpuslinguistische Verfahren bezeichnet, bei denen aufgrund von zuvor aufgestellten Hypothesen über sprachliche Verknüpfungen die Verteilung eines im Voraus definierten lexikalischen Elements in einem definierten Teilkorpus (bspw. in einem bestimmten Zeitabschnitt oder in den Texten einer bestimmten Sprecherposition) untersucht wird. Als *corpus driven* werden hingegen induktive Verfahren bezeichnet, die ohne im Voraus definierte Suchanfragen auskommen und damit die Chance bieten, auf Strukturen zu stoßen, an die man nicht schon vor der Untersuchung gedacht hat (Tognini-Bonelli 2001). Ein *corpus driven*-Vorgehen ist daher besonders für explorative Zwecke geeignet, d. h. um einen ersten Überblick über Unterschiede und Gemeinsamkeiten sprachlicher Verweisstrukturen aufzuzeigen.

Folgt man der gemeinsamen theoretischen Grundannahme von Strukturalismus und Poststrukturalismus, dass Bedeutung ein Effekt der Beziehung von (lexikalischen) Elementen zu anderen (lexikalischen) Elementen ist, dann können lexikometrische Verfahren herangezogen werden, um diese Beziehungen und damit die Konstitution von Bedeutung herauszuarbeiten.[3] Bislang werden lexikometrische Verfahren al-

3 John Rupert Firth hat diese Perspektive als „Kontextualismus" bereits in den 1950er-Jahren in die Linguistik eingeführt. Die Kontexte lexikalischer Elemente geben danach Hinweise auf deren Gebrauch und damit deren

lerdings kaum im Rahmen von Forschungsprojekten eingesetzt, die auf eine Operationalisierung von Diskurstheorien – etwa im Anschluss an Foucault oder Laclau und Mouffe – zielen. Wo liegen die Ursachen für diese Zurückhaltung? Baker (2006) vermutet, dass lexikometrische Verfahren als quantitative Methoden vielfach in eine Schublade mit Ansätzen gesteckt werden, die in einem naiven Realismus davon ausgehen, dass wissenschaftliche Analysen einfach objektive Fakten messen können (2006: 8). Hinzu kommt, dass lexikometrische Verfahren vielfach selbst bei diskursanalytisch arbeitenden Sozialwissenschaftler_innen nicht bekannt sind. Auch in sozial- und kulturwissenschaftlich orientierten Bereichen der deutsch- und englischsprachigen Sprachwissenschaften werden bislang nur vereinzelt korpusbasierte lexikometrische Verfahren angewendet (für die deutschsprachige Diskursgeschichte s. aber bspw. Jung 1994 und Niehr 1999, für die – überwiegend englischsprachige – *critical discourse analysis* s. aber bspw. Orpin 2005 sowie Baker, Gabrielatos, Khosravinik, Kryzanowski, Mcenery und Wodak 2008).[4] In der Konsequenz bleibt der Einsatz lexikometrischer Verfahren vielfach auf Bereiche der Linguistik beschränkt, die kaum im Austausch mit gesellschaftstheoretischen Überlegungen stehen. Dabei werden lexikometrische Verfahren teilweise in Forschungsdesigns eingebunden, die in einer rein strukturalistischen Perspektive darauf abzielen, sprachliche Strukturen zu messen (kritisch dazu bspw. der Sprachwissenschaftler Teubert 1999).[5]

Lexikometrische Verfahren können jedoch durchaus sinnvoll in ein Forschungsdesign eingebunden werden, das entsprechend einer diskurs-

---

Bedeutung: *„You shall know a word by the company it keeps“* (1957, zit. nach Belica und Steyer 2005; Lemnitzer und Zinsmeister 2006).

4 Anders stellt sich die Situation in der französischsprachigen Forschungslandschaft dar: Hier existieren im Rahmen der so genannten „französischen Schule der Diskursforschung“ seit den 1960er-Jahren vielfältige Beziehungen zwischen Linguistik, Politik- und Geschichtswissenschaft, so dass politik- und geschichtswissenschaftliche Arbeiten regelmäßig auch auf lexikometrische Verfahren zurückgreifen (Bonnafous und Tournier 1995; Guilhaumou 1997; Mayaffre 2004). In Deutschland wurden diese Arbeiten bislang nur vereinzelt von einigen Romanisten rezipiert (Lüsebrink 1998; Reichardt 1998), einen englischsprachigen Überblick liefert Williams 1999. Einige Hinweise zur Verwendung lexikometrischer Verfahren in der französischsprachigen Geschichtswissenschaft bietet auch der von Reiner Keller ins Deutsche übersetzte Aufsatz von Guilhaumou 2003.

5 Dies scheint wiederum Wissenschaftler, die an der Diskurstheorie von Laclau und Mouffe anknüpfen, in ihren Vorbehalten gegenüber Verfahren zu bekräftigen, die Marchart bspw. pauschal als „rein statistisches Wörterzählen“ ablehnt (1998: FN 19).

theoretischen Positionierung auf die Kontingenz und Dynamik von Bedeutungen abhebt (ähnlich argumentieren bspw. Teubert 1999, 2005; Koteyko 2006; Glasze 2007b; Mattissek 2008)[6].

Exkurs: Abgrenzung der Lexikometrie von der quantitativen Inhaltsanalyse

Die Verfahren der Lexikometrie sind nicht zu verwechseln mit Verfahren der quantitativen Inhaltsanalyse. Die Inhaltsanalyse hat ihre Ursprünge in der US-amerikanischen Kommunikationswissenschaft und baut auf einem Kommunikationsmodell „Sender ➔ Inhalt ➔ Empfänger" auf. Ziel ist dabei, vom „Inhalt" auf „die soziale Wirklichkeit" zu schließen – d. h. auf den „Kommunikator", den „Rezipienten" oder die „Situation". Wichtigste Methodik ist die Kodierung der „Inhalte" von Texten mittels eines Kategoriensystems (Merten 1995). Die Inhaltsanalyse geht dabei davon aus, dass jeder Text(-teil) einen eindeutig zu bestimmenden „Inhalt", d. h. *eine* Bedeutung, transportiert und dass dieser „Inhalt" durch die Inhaltsanalytiker_innen erschlossen werden kann. Aus der Sicht einer poststrukturalistisch informierten Diskursforschung ist ein solches Repräsentationsmodell, das von einem Text(-teil) auf *die* eine, vermeintlich gegebene Bedeutung schließen will, problematisch. Die Diskursforschung betont ja gerade die Mehrdeutigkeit und Instabilität von „Sinn" (vgl. Kap. 1). Insbesondere französische Diskursforscher kritisierten die Diskursforschung dafür, dass die Arbeit mit Kategoriensystemen zudem das Risiko mit sich bringt, Tautologien zu erzeugen, indem ein voretabliertes System durch Belegstellen reifiziert wird (s. Bonnafous 2002).

Berelson, der als einer der Väter und Vordenker der quantitativen Inhaltsanalyse bezeichnet werden kann, plädierte allerdings Anfang des 20. Jahrhunderts dezidiert für eine Inhaltsanalyse, die am „manifesten Inhalt" und an den *„black-marks-on-white"* ansetzt (Berelson 1952: 19). Tatsächlich gehen Frequenzanalysen von Wörtern und Wortfolgen, wie sie in quantitativen Inhaltsanalysen in der Tradition von Berelson durchgeführt werden, ähnlich vor wie Frequenzanalysen in der Lexikometrie bzw. Korpuslinguistik. Der theoretische Hintergrund und damit der Stellenwert, der den Ergebnissen zugespro-

6 Die Lexikometrie ermöglicht es bspw., grundlegende Prinzipien der Diskurstheorie nach Laclau und Mouffe (Kap. 6) zu operationalisieren: So werden die „Elemente" der Diskurstheorie als lexikalische Formen gefasst, die in temporären Fixierungen zu „Momenten" eines Diskurses werden. Das Konzept der „Regelmäßigkeit von Differenzbeziehungen" wird operationalisiert als die mit einer gewissen Signifikanz verknüpften lexikalischen Elemente. Die Signifikanz ist dabei das Maß der Überwahrscheinlichkeit für das Auftreten eines lexikalischen Elements im Kontext eines anderen Elements (s. Kap. 2 und Exkurs „Lexikometrische bzw. korpuslinguistische Software und Ressourcen"). Dabei kann die Temporalität jeglicher Fixierung mittels einer vergleichenden Untersuchung verschiedener (Sub-)Korpora im Zeitvergleich herausgearbeitet werden (Glasze 2007b).

chen wird, ist jedoch ein anderer: In der Inhaltsanalyse werden lexikalische Elemente wie einzelne Wörter nicht wie in der Lexikometrie als „Bausteine" der Konstitution von Bedeutung, sondern unmittelbar als Indikatoren für die „soziale Wirklichkeit" interpretiert, indem ihnen eine denotative „Standardbedeutung" zugeschrieben wird (Merten 1995). Berelson war sich zwar der engen Grenzen einer solchen Perspektive durchaus bewusst. Sein Lösungsvorschlag lautet jedoch etwas naiv: „…content analysis must deal with relatively denotative communication materials and not with relativeley connotative materials". Er nennt „Nachrichtenmeldungen" *(news stories)* als Beispiel für „denotative Kommunikationen" (Berelson 1952: 19f.). Auch neuere computergestützte Verfahren der Inhaltsanalyse gehen von der Prämisse aus, dass die Bedeutung von Wörtern feststeht. Sie arbeiten mit „a priori Wörterbüchern", welche Wörter „gleicher Bedeutung" auflisten, die von den Programmen gemeinsam kategorisiert werden (Atteslander und Cromm 2006: 202ff.). Wie insbesondere Roland Barthes gezeigt hat, kann allerdings eine Grenze zwischen der einen denotativen „Standardbedeutung" und weiteren konnotativen Nebenbedeutungen nicht sinnvoll gezogen werden (s. Kap. 1).

Innerhalb der Sozialforschung wurde versucht, den skizzierten Problemen der Inhaltsanalyse insofern zu begegnen, als Methoden entwickelt wurden, welche die Interpretation methodisch kontrollieren und intersubjektiv absichern sollen. In diese Richtung zielen bspw. die Vorschläge einer „Objektiven Hermeneutik" nach Oevermann (Wernet 2006) oder auch der „Qualitativen Inhaltsanalyse" nach Mayring (2008 [1983]). In den vergangenen Jahren hat zudem ein Austausch zwischen Inhaltsanalyse und Sprachwissenschaften eingesetzt und in der Forschungspraxis finden sich Überschneidungen von Inhaltsanalysen und sprachwissenschaftlichen Verfahren. Die Unterschiede in der theoretischen Fundierung bleiben jedoch bestehen: Aus einer poststrukturalistisch informierten Perspektive fehlt der Inhaltsanalyse nach wie vor eine Auseinandersetzung mit einer Theorie der Bedeutungskonstitution (so auch der Inhaltsanalytiker Merten 1995).

Verfahren der Lexikometrie und Korpuslinguistik

Auswahl der relevanten Textsorten

Grundlage lexikometrischen Arbeitens sind digitale Textkorpora.[7] In den Analysen werden unterschiedliche Teile des Korpus miteinander verglichen. Korpora für lexikometrische Analysen müssen „geschlossen" sein, da die lexikometrischen Analysen nur dann sinnvoll sind, wenn sie sich auf ein stabiles Ensemble von Texten beziehen. Für die Zusammenstellung des Korpus ist es entscheidend, dass – mit Ausnahme der zu analy-

7 Dafür wird entweder auf Texte zurückgegriffen, die bereits digital vorliegen, oder die Texte müssen mittels Texterkennung digitalisiert werden.

sierenden Variable (bspw. unterschiedliche Zeitabschnitte oder unterschiedliche Sprecherpositionen) – die Bedingungen der Aussagenproduktion möglichst stabil gehalten werden (s. u.). Denn bei einem Vergleich, bei dem zwischen den zu vergleichenden Teilen sowohl die Zeit bzw. Epoche, die Kommunikationskanäle, die Sprecherposition, die Genres etc. wechseln, könnten keine sinnvollen Ergebnisse gewonnen werden, da nicht mehr bestimmt werden kann, auf welche Veränderungen sprachliche Unterschiede zurückzuführen sind.

Bei der Vorbereitung lexikometrischer Verfahren ist die Überlegung zentral, bezüglich welcher Kriterien die Bedeutungskonstitution verglichen werden soll, denn dies entscheidet über die Segmentierung, d. h. Aufteilung des Textkorpus in entsprechende vergleichbare Teilkorpora: Sollen zeitliche Verschiebungen untersucht werden, wird man eine diachrone Segmentierung wählen. Geht es darum, die Unterschiedlichkeit von Bedeutungskonstitutionen aus der Sicht von einzelnen Sprecherpositionen oder in einzelnen Genres zu erfassen, wird man einen zeitlich homogenen Korpus wählen, der nach den auftretenden Sprecherpositionen bzw. nach einzelnen Genres segmentiert wird etc. Die anderen Merkmale des Textes werden dabei jeweils konstant gehalten.[8]

Sprecherpositionen werden hier in Anlehnung an Überlegungen Foucaults (1973 [1969]) als institutionell stabilisierte Positionen i. d. R. innerhalb von Organisationen gefasst, die spezifische Zugangskriterien haben und die bestimmte Möglichkeiten, Tabus und Erwartungen des Sprechens bzw. allgemein der Textproduktion mit sich bringen – weitgehend unabhängig von den Individuen, welche die Position einnehmen. Die Sprecherpositionen sind dabei selbst diskursiv konstituiert. Je nach Fragestellung der Untersuchung kommen unterschiedliche Sprecherpositionen infrage. In der Regel sind gesellschaftlich bedeutsame Sprecherpositionen in Organisationen eingebunden, sind also Positionen, von denen aus *im Namen* und *als* Organisation gesprochen werden kann (bspw. Texte von Wissenschaftsorganisationen, Zeitungstexte, Texte von Behörden etc.). Für diachron angelegte Studien ist darüber hinaus die Arbeit mit Serien sinnvoll, die durch regelmäßige Publikationen einer Sprecherposition entstehen (Texte regelmäßig erscheinender Medien, Verhandlungsbände regelmäßig stattfindender Konferenzen, Protokolle regelmäßig tagender Gremien etc.).

Mit dem Begriff des Genre bzw. der Gattung werden in den Sprachwissenschaften Gruppen von Texten bezeichnet, für deren Strukturie-

8 Die Zusammenstellung des Korpus ist also abhängig von der Fragestellung der Untersuchung, wobei immer auch die Frage geklärt werden muss, wofür ein bestimmter Korpus steht.

rung und damit deren Kohärenz sich historisch spezifische, institutionell stabilisierte Regeln etabliert haben (Maingueneau 2000 [1986]): So gelten für die Strukturierung und Kohärenz wissenschaftlicher Fachaufsätze andere Regeln als für Zeitungsartikel und wiederum andere für politische Reden. Im Rahmen der Diskurstheorie können die entsprechenden Institutionen, d. h. Sprecherpositionen bzw. Genres, selbst als diskursiv konstituiert konzeptionalisiert werden – als „sedimentierte Diskurse" (Laclau 1990 und s. Kap. 6 zur Diskurs- und Hegemonietheorie).[9]

Zusammenstellung der Texte für das Textkorpus

Für die Zusammenstellung der Texte für das zu untersuchende Textkorpus lassen sich zwei prinzipielle Strategien unterscheiden:

1. Es werden alle Texte einer bestimmten Textserie (z. B. alle Texte einer bestimmten Zeitung, alle Reden eines Präsidenten, alle Verlautbarungen einer Organisation etc.) über einen festen Zeitraum berücksichtigt.
2. Anhand des Auftretens von Schlüsselwörtern oder thematischen Kodierungen wird ein thematisches Korpus erstellt.

Das zweitgenannte Verfahren erscheint dabei insofern problematisch, als ein solches Vorgehen immer Gefahr läuft, dass nur jene Texte bzw. Textpassagen berücksichtigt werden, die den impliziten Erwartungen der Wissenschaftler entsprechen (Baker 2006). Eine Arbeit mit thematisch zusammengestellten Textkopora, wie sie Busse und Teubert (1994: 14) vorgeschlagen haben, scheint daher für lexikometrische Studien nicht geeignet.[10] Darüber hinaus kann argumentiert werden, dass nicht alle sprachlichen Muster an Schlüsselbegriffen festzumachen sind: So kön-

9 In den Sprachwissenschaften gibt es zudem Bemühungen, durch die Zusammenstellung sehr großer Textmengen Standardkorpora zu erstellen, welche „den" typischen Sprachgebrauch einer bestimmten Epoche abbilden sollen. Beispiele sind der *British National Corpus* (BNC) (verfügbar unter http://www.natcorp.ox.ac.uk, Zugriff: 25.9.2006), das Projekt Digitales Wörterbuch der deutschen Sprache (verfügbar unter http://www.dwds.de, Zugriff: 25.9.2006), die Korpora des Instituts für Deutsche Sprache in Mannheim (verfügbar unter http://www.ids.de, Zugriff: 16.1.2007) sowie der französische Korpus *Frantext* (verfügbar unter http://www.frantext.fr, Zugriff: 25.9.2006). Ziel solcher Bemühungen ist es u. a., Vergleiche zwischen spezifischen Textkorpora und „dem" Sprachgebrauch in einer Epoche zu ermöglichen.

10 Für die stärker interpretativ ausgerichteten Methoden, wie bspw. kodierende Verfahren (s. Kap. 14), kann hingegen mit offenen Korpora gearbeitet werden, die im Laufe der Analyse verändert, d. h. verkleinert bzw. ergänzt, werden. Dann kann auch mit einem auf Basis des Kontextwissens thematisch zusammengestellten Korpus begonnen werden.

nen etwa oft als „neoliberal“ bezeichnete sprachliche Formen auf ganz unterschiedliche Art und Weise und mithilfe sehr unterschiedlicher Wortverbindungen ausgedrückt werden (Mattissek 2008). Die Analyse geschlossener Korpora, bspw. mit Serien von Texten einer homogenen Sprecherposition, begrenzt das Risiko von Zirkelschlüssen und erhöht die Chance, Diskursmuster herausarbeiten zu können, die *nicht* den impliziten Erwartungen entsprechen (Glasze 2007b).

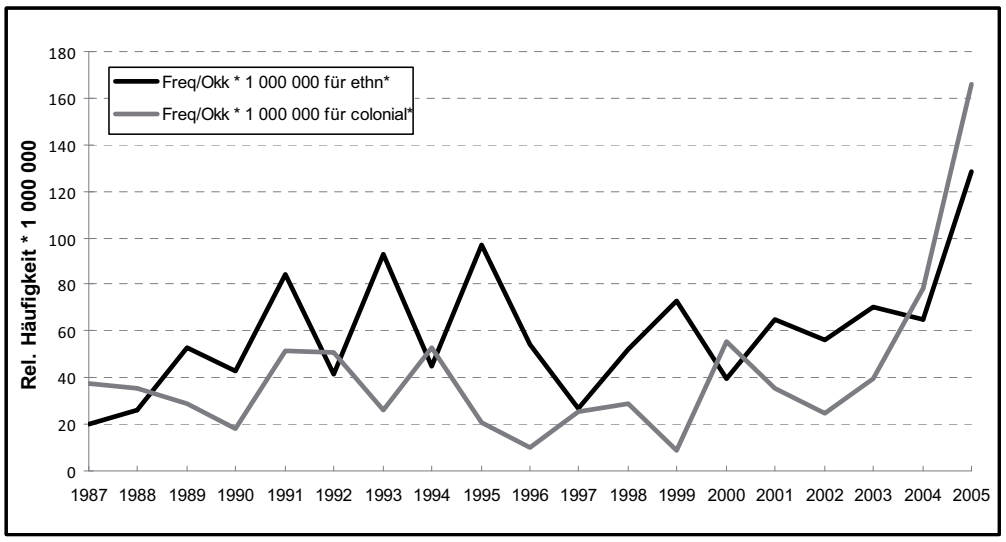
Verfahren der lexikometrischen Analyse von Texten

Innerhalb lexikometrischer Verfahren lassen sich insbesondere vier grundlegende Methoden unterscheiden, die vielversprechende Ansätze für diskurstheoretisch orientierte Arbeiten bieten. Dies sind Frequenz- und Konkordanzanalysen, Analysen der Charakteristika eines Teilkorpus sowie Analysen von Kookkurrenzen (eine ausführlichere Darstellung bieten Lebart und Salem 1994; Lebart, Salem und Berry 1998; Marchand 1998: 29ff.; Baker 2006). Für die Berechnung bzw. Erstellung der jeweiligen Parameter und Auswertungen kann auf unterschiedliche Computerprogramme zurückgegriffen werden (s. Exkurs). Im Folgenden werden die wichtigsten lexikometrischen Verfahren anhand empirischer Beispiele kurz vorgestellt.

1. *Frequenzanalysen* zeigen, wie absolut oder relativ häufig eine spezifische Form in einem bestimmten Segment des Korpus auftritt: Auf der Basis von diachronen Korpora lassen sich damit also bspw. die relative Häufigkeit eines Wortes (Graphems, d. h. der kleinsten zusammenhängenden Einheit, die im Schreibfluss auftritt) oder von regelmäßig verknüpften Wörtern (Wortfolgen, bzw. N-Grammen) im Zeitverlauf herausarbeiten (vgl. Abbildung 4). In der Regel werden zunächst Wortlisten erstellt, in denen alle grammatischen Formen eines Wortes (d. h. die einzelnen Grapheme) getrennt gezählt werden (also bspw. Forscher/Forscherin/Forschern/Forschers). Oftmals kommt es aber vor, dass man solche Flexionen für die Analyse als gleichwertig betrachten möchte. In diesem Fall wird mit dem *Lemma* bzw. *Lexem* gearbeitet, d. h. mit einer Gruppe verschiedener Flexionsformen, die alle zum gleichen Begriff gehören; im o. g. Beispiel würden also Forscher/Forschern und Forschers zum gleichen Lemma gehören. Dieser Vorgang wird als *Lemmatisierung* bezeichnet (vgl. Lebart, Salem und Berry 1998: 23). Die Grenze, welche Wortformen für eine gegebene Analyse als äquivalent angesehen werden und welche nicht, hängt von der Fragestellung ab – so kann etwa in einem Fall, wo es um ungleiche Geschlechterverhältnisse an Universitäten geht, die Unterscheidung zwischen „Forscher“ und „Forsche-

rin“ entscheidend sein, in einem anderen Fall, wo es nur darum geht, welche Rolle in einer bestimmten Stadt die Forschung spielt, ist diese Differenzierung nicht relevant (vgl. Lebart, Salem und Berry 1998: 22).

Abbildung 4: Ergebnisse des diachronen Vergleichs relativer Häufigkeiten von Wörtern des Postkolonialismus und der ethnischen Differenzierung im Banlieue-Diskurs (Le Monde 1987–2005)



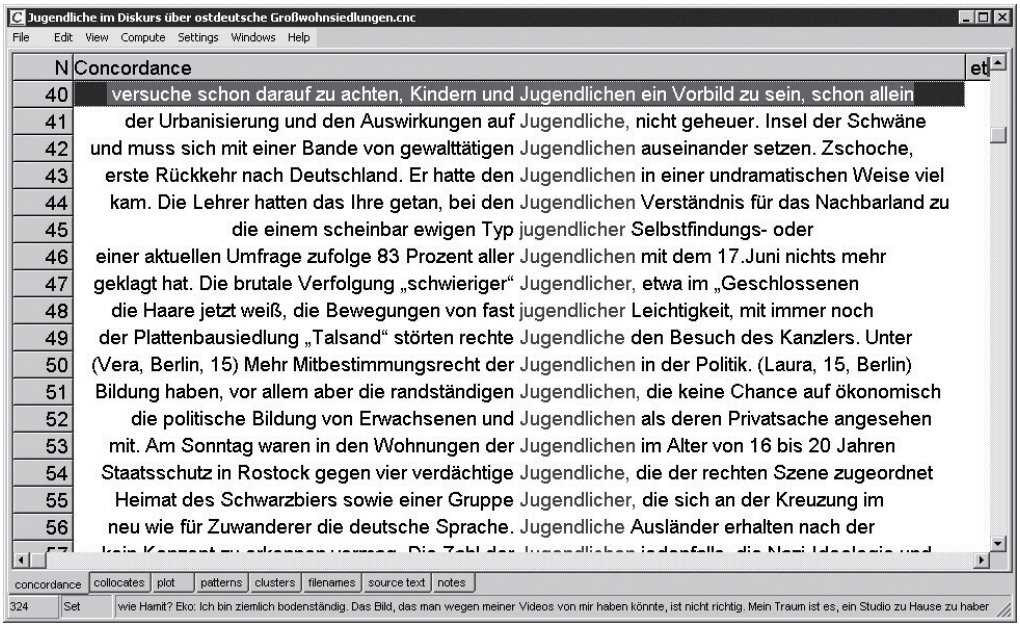
Quelle: eigene Darstellung

Die Abbildung zeigt ein Teilergebnis eines Forschungsprojekts, in dem der Bedeutungswandel des Lexems „banlieue*“ im Kontext der so genannten „Krise der *banlieues*“ in Frankreich untersucht wird (Glasze, Germes und Weber 2009). Anhand eines diachron angelegten Pressekorpus mit allen Artikeln der Tageszeitung Le Monde seit 1987, die das Lexem „banlieue“ enthalten, konnte gezeigt werden, dass seit 2003 die relative Häufigkeit von Wörtern stark angestiegen ist, die semantisch dem Feld der ethnischen Differenzierung (untersucht wurde die Buchstabenfolge ethn*) und dem Postkolonialismus (untersucht wurde die Buchstabenfolge colonial*) zugerechnet werden können[11]. Dies kann als ein erster Hinweis darauf gewertet werden, dass es zu einer Ethnisierung des *banlieue*-Diskurses kommt, d. h. dass die so genannte „Krise der *banlieues*“ seit 2003 zunehmend als eine Krise der ethnischen Differenz und des Postkolonialismus konstituiert wird.

11 Die Lexeme ethn* und colonial* wurden ausgewählt, da diese in einer vorangegangenen *Analyse der Charakteristika eines Teilkorpus* (s. unten) als Charakteristika des *banlieue*-Diskurses identifiziert wurden.

2. *Konkordanzanalysen*. Die einfachste Möglichkeit, den Kontext eines Wortes bzw. einer Wortfolge zu untersuchen, ist die Anzeige von Konkordanzen. Dabei werden die jeweils vor und hinter einem Schlüsselwort stehenden Zeichenfolgen dargestellt. Eine Konkordanz ist eine Liste, die alle Vorkommen eines ausgewählten Wortes – oder auch Wortfolgen – in seinem Kontext zeigt. Für Konkordanzen üblich ist eine zeilenweise Darstellung, die als KWIC *(key word in context)* bezeichnet wird (Scherer 2006: 43; vgl. Lebart, Salem und Berry 1998: 32f.). Auf dessen linker und rechter Seite wird, je nach verwendeter Analysesoftware, ein festgelegter Kontext, bestehend aus einer bestimmten Anzahl an Zeichen oder Wörtern, angezeigt (vgl. Abbildung 5). Konkordanzanalysen können sinnvoll als Vorbereitung und Hilfe für die qualitative Interpretation des Kontextes bestimmter Schlüsselwörter verwendet werden.

Abbildung 5: Ausschnitt aus einer Konkordanzliste



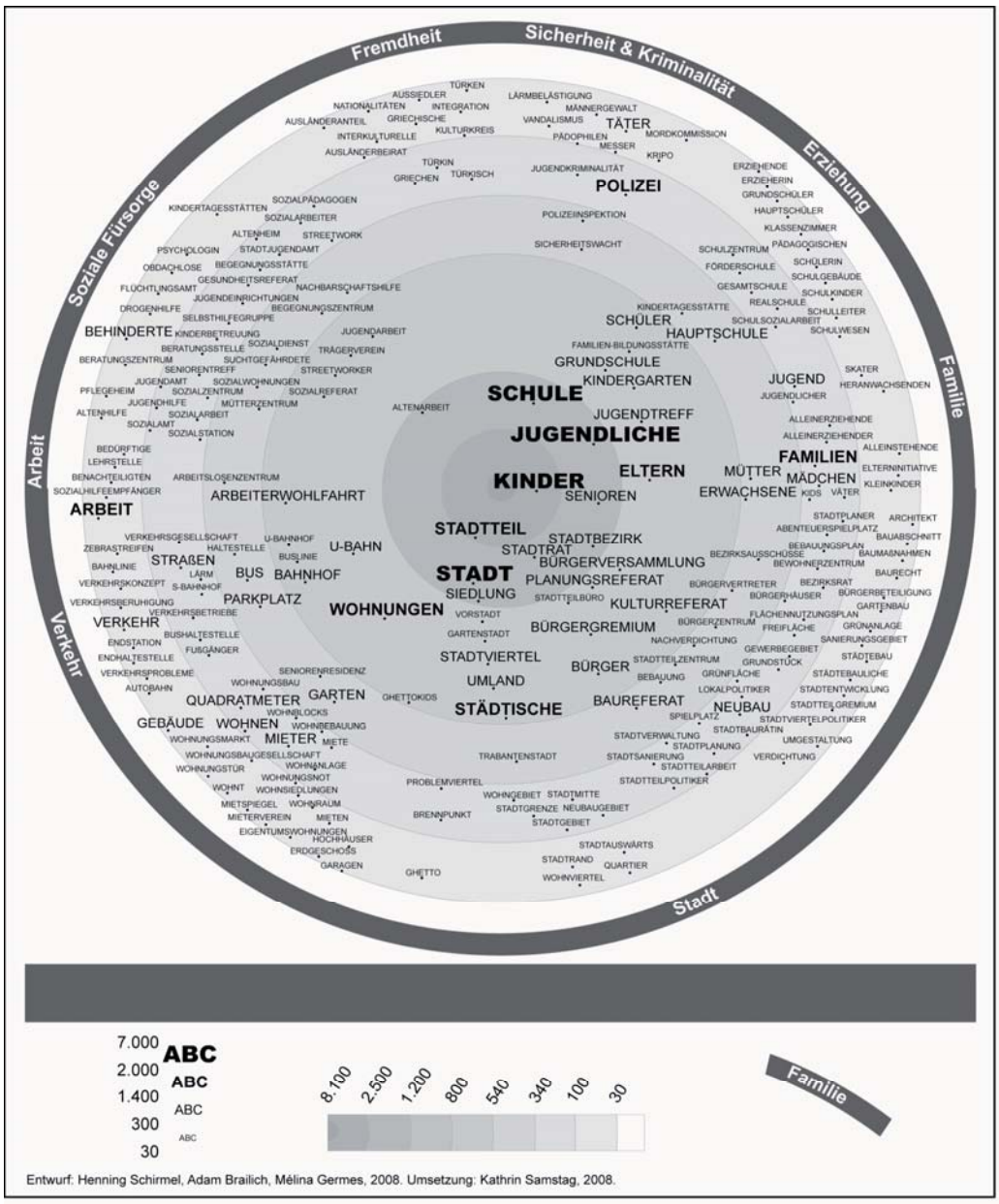
Quelle: eigene Darstellung

Die Abbildung zeigt einen Ausschnitt der Konkordanzen des Lexems „Jugendliche*“ in dem Diskurs über die ostdeutschen Großwohnsiedlungen der Süddeutschen Zeitung von 1994–2006 (Brailich, Germes, Glasze, Pütz und Schirmel 2009). Das Lexem „Jugendliche“ stellt in der hier gezeigten Studie über die Konstitution ostdeutscher Großwohnsiedlungen ein Charakteristikum eines Teilkorpus dar und wird anhand einer Konkordanzanalyse kontextualisiert. So lässt sich in der Darstellung der Konkordanzen erkennen, dass „Jugendliche*“ vielfach in Verbindung mit Wörtern steht, welche Problematisierungen als „gewalttätige“, „rechtsextreme“ oder „randständische Jugendliche“ konstituieren.

3. Analysen der *Charakteristika eines Teilkorpus* zeigen, welche lexikalischen Formen für einen Teil des Korpus im Vergleich zum Gesamtkorpus bzw. einem anderen Teilkorpus spezifisch sind. Hierzu werden diejenigen Wörter ermittelt, die in einem bestimmten Teilkorpus signifikant über- oder unterrepräsentiert sind[12]. Die Analysen von Charakteristika eines Teilkorpus sind also induktiv und *corpus driven*, d. h. sie kommen ohne im Voraus definierte Suchanfragen aus und bieten damit die Chance, auf Strukturen zu stoßen, an die man nicht schon vor der Untersuchung gedacht hat (Teubert 2005; Bubenhofer 2008).

12 Grundlage der Berechnung der Signifikanz sind die absolute Häufigkeit eines bestimmten Wortes bzw. einer Gruppe von Wörtern (d. h. von Graphemen oder Lexemen) bzw. einer Wortfolge und die Gesamtzahl aller Wörter in einem gegebenen Korpus (Okkurrenzen). Aus dem Verhältnis zwischen der Häufigkeit einzelner Wörter bzw. Wortgruppen und der Gesamtzahl aller Wörter im Korpus lässt sich die Wahrscheinlichkeit für eine bestimmte Frequenz des Wortes in einem Teil des Korpus berechnen. Auf diese Weise lässt sich zeigen, welche Wörter und ggf. Wortfolgen in einem Teilkorpus im Vergleich zum Gesamtkorpus spezifisch häufiger bzw. seltener vorkommen. Je nach Analysesoftware stehen unterschiedliche statistische Tests für die Berechnung der Signifikanzen zur Verfügung (s. Exkurs „Lexikometrische bzw. korpuslinguistische Software und Ressourcen“).

Abbildung 6: Ergebnisse der Analyse von Charakteristika des Teilkorpus westdeutsche Großwohnsiedlungen mit dem Referenzkorpus Süddeutsche Zeitung 1994–2006



Quelle: eigene Darstellung

Die Abbildung zeigt die in einer Analyse der Süddeutschen Zeitung (1994–2006) herausgearbeiteten Charakteristika eines Teilkorpus und visualisiert den Mediendiskurs zu den westdeutschen Großwohnsiedlungen (Brailich, Germes, Glasze, Pütz und Schirmel 2009). Dazu wurde anhand der Eigennamen der westdeutschen Großwohnsiedlungen mit mehr als 2.500 Wohneinheiten ein Teilkorpus gebildet und mit dem Referenzkorpus aller Artikel der Süddeutschen Zeitung verglichen. Mit diesem *corpus driven*-Ansatz wurden diejenigen Wörter herausgearbeitet, die charakteristisch für die Bedeutungskonstitution westdeutscher Großwohnsiedlungen sind. Je näher die charakteristischen

Wörter zum Zentrum der Abbildung stehen, desto signifikanter sind diese (desto höher ist ihr keyness-Wert[13]). Der Schriftgrad zeigt die Häufigkeit des jeweiligen Wortes im Teilkorpus an. Die Wörter wurden „manuell" in thematischen Gruppierungen angeordnet, um eine bessere Übersichtlichkeit zu gewährleisten.[14]

4. Die Untersuchung von *Kookkurrenzen*[15] zeigt, welche Wörter und Wortfolgen (N-Gramme) im Korpus mit einer gewissen Signifikanz miteinander verknüpft werden, d. h. welche Wörter in der Umgebung eines bestimmten Wortes überzufällig häufig auftauchen[16]. Dafür wird ein Teilkorpus mit der Umgebung um ein bestimmtes Schlüsselwort erstellt. Diese Umgebungen können Einheiten sein wie der Satz oder Absatz, in dem das Schlüsselwort vorkommt, ein definierter Bereich mit einer bestimmten Zahl von Wörtern vor und nach dem Schlüsselwort oder Einheiten, aus denen der Korpus zusammengesetzt wurde (bspw. einzelne Reden oder Presseartikel, in denen das Schlüsselwort vorkommt). Der Teilkorpus mit den Wörtern und Wortfolgen in der Umgebung des Schlüsselwortes wird dann auf Charakteristika im Vergleich zum Gesamtkorpus untersucht (s. o.).

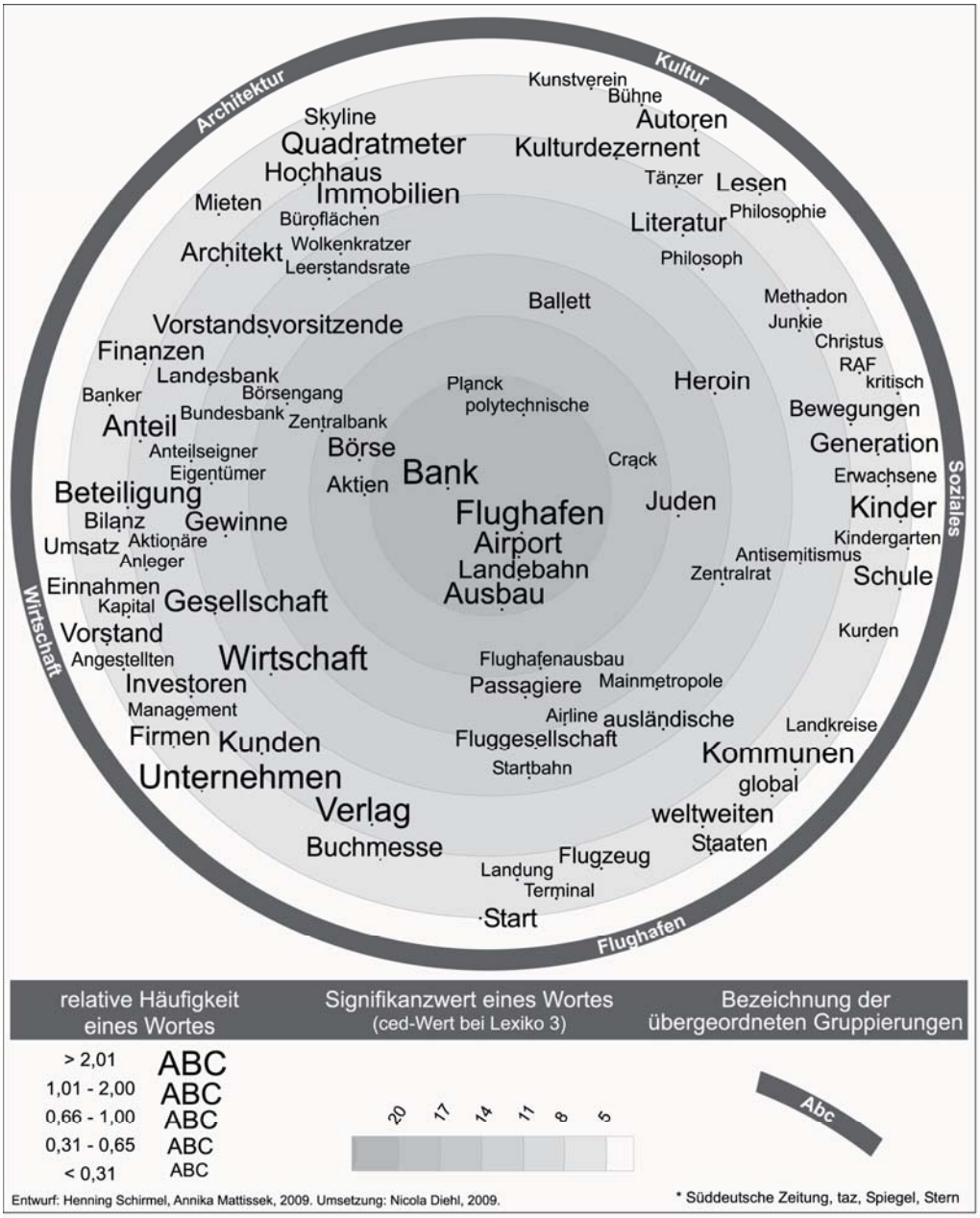
13 Das lexikometrische Analyseprogramm WordSmith Tools berechnet als Signifikanzwert der Charakteristika einen so genannten keyness-Wert. Die Berechnung der Charakteristika erfolgte mithilfe eines statistischen Tests auf Basis von Ted Dunnings Log Likelihood (Dunning 1993).

14 Als Interpretationshilfe wurden dabei Konkordanzanalysen herangezogen, um wiederum den Kontext der Charakteristika genauer bestimmen zu können (s. o.).

15 Teilweise werden Wörter, die regelmäßig in der Nähe voneinander auftreten, auch als Kollokationen bezeichnet (Baker 2006).

16 Zur Berechnung von Signifikanzen s. unter (3.) „Analyse der Charakteristika eines Teilkorpus".

Abbildung 7: Ergebnisse der Kookkurrenzanalyse des Begriffs „Frankfurt" in überregionalen Printmedien 1999–2005*



Quelle: eigene Darstellung

Die Abbildung zeigt die in der vergleichenden Printmedienanalyse herausgearbeiteten Kookkurrenzen mit „Frankfurt am Main" (Mattissek 2008). Je weiter innen die Begriffe stehen, desto höher der ced-Wert[17] und desto wahrscheinlicher ist es folglich, dass die beobachtete Häufung dieser Begriffe in den Artikeln zu Frankfurt statistisch signifikant (d. h. „überzufällig") ist. Die Größe der Begriffe entspricht der relativen Häufigkeit im untersuchten Teilkorpus.

17 Der ced-Wert ist ein Maß für die Überzufälligkeit des Auftretens eines Begriffs in einem Teilkorpus verglichen mit allen anderen Teilkorpora.

Inhaltlich-semantisch wurden die Wörter der Übersichtlichkeit halber manuell unterschiedlichen thematischen Segmenten zugeordnet.

5. Eine sinnvolle Erweiterung der Kookkurrenzanalyse bieten *multivariate Analyseverfahren* von Differenzbeziehungen, mithilfe derer sich Kookkurrenzen verschiedener Begriffe in unterschiedlichen Teilkorpora in einen Zusammenhang bringen lassen (Dzudzek 2008). Während bei der Kookkurrenzanalyse die Bedeutungsverschiebung *eines* Begriffs durch unterschiedliche Teilkorpora verfolgt wird und man davon ausgeht, dass es einen fixen Knotenpunkt gibt, der in allen Teilkorpora eine zentrale, wenn auch sich verändernde Rolle spielt, bieten multivariate Analyseverfahren von Differenzbeziehungen die Möglichkeit, diskursive Verschiebungen im Sinne eines Gleitens von „Signifikant zu Signifikant" zwischen unterschiedlichen Teilkorpora zu verfolgen. Ziel multivariater Analyseverfahren von Differenzbeziehungen ist also die Erweiterung der Kookkurrenzanalyse hin zu einer Analyse eines komplexen differenziellen Netzes von Zeichen (Textelementen), in dem Bedeutung konstituiert wird im Sinne von de Saussure (1931 [1916]) sowie Laclau und Mouffe (1985). Daher wird der Begriff Kookkurrenz, der „das gemeinsame Vorkommen *zweier* Wörter in einem gemeinsamen Kontext" (Lemnitzer und Zinsmeister 2006: 147) beschreibt, um eine Vielzahl von Dimensionen erweitert, so dass er die Beziehung *mehrerer* Elemente innerhalb eines differenziellen Zeichensystems beschreibt. Mithilfe multivariater Dimensionsreduktionsverfahren lässt sich dieser komplexe n-dimensionale Zusammenhang in einem zweidimensionalen Koordinatensystem visualisieren. Als dimensionsreduzierende Methoden für linguistische Sachverhalte kommen die Hauptkomponentenanalyse und die Korrespondenzanalyse[18] infrage (vgl. Lebart, Salem und Berry 1998: 45–69). In diesem zweidimensionalen Koordinatensystem lassen sich nun folgende Sachverhalte ablesen:
   a) Möglicher *Kontext*: Die relative Nähe bzw. Entfernung der Begriffe zueinander ist Indikator für mögliche Differenz- bzw. Äquivalenzverhältnisse im Sinne von Laclau und Mouffe.
   b) Möglicher relativer *Bedeutungsgrad* für den Diskurs: Je weiter die Begriffe vom Mittelpunkt des Koordinatenkreuzes entfernt

18 Es gibt eine Vielzahl von Statistikprogrammen, die unterschiedliche multivariate Verfahren rechnen und visualisieren können. Das – eigentlich für die Ökologie entwickelte – Programm Canoco 4.5 kann neben der üblichen Hauptkomponentenanalyse auch die für linguistische Zwecke entwickelte Korrespondenzanalyse rechnen (Ter Braak und Smilauer 2002).

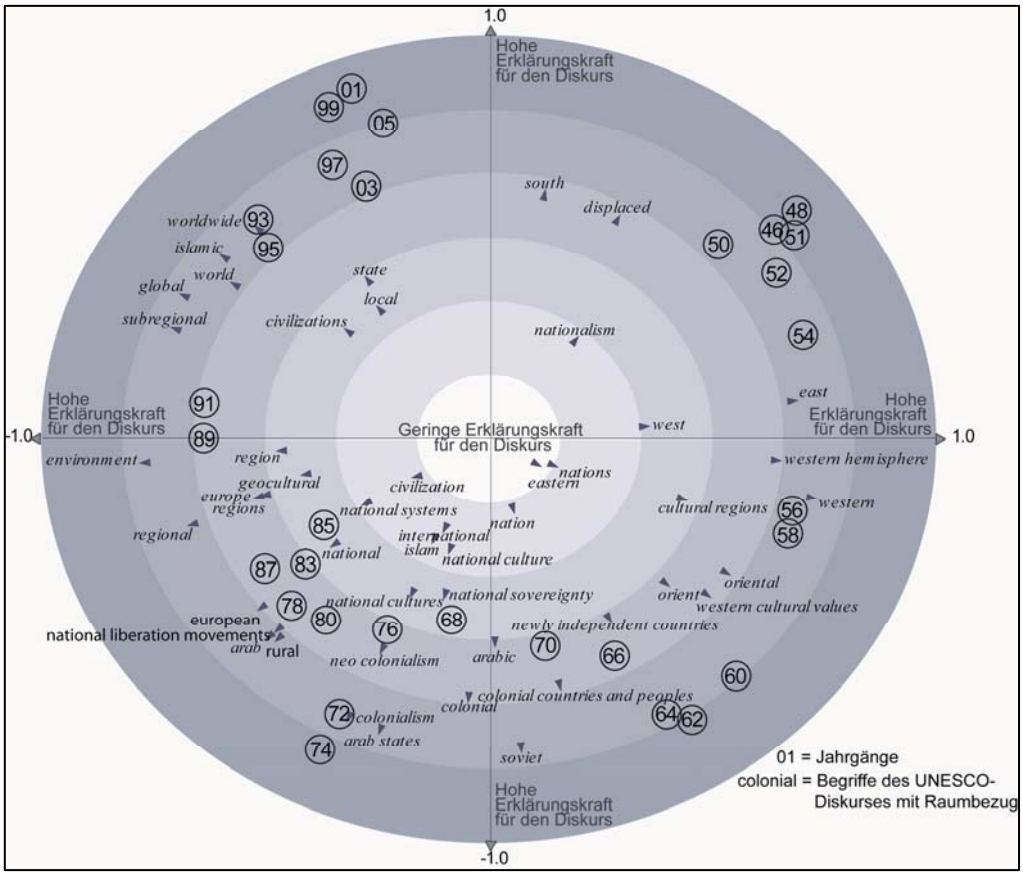
sind, desto größer ist ihre „relative Wichtigkeit“ (Lebart, Salem und Berry 1998: 57) für den Diskurs einzuschätzen. Denn je häufiger ein Begriff in einem Teilkorpus vorkommt und je stärker er sich vom Vorkommen in sonstigen Teilkorpora und von anderen Begriffen unterscheidet, desto stärker wird er nach außen gezogen.

c) Mögliche *Cluster von Teilkorpora*: Nicht nur die Lage der Begriffe zueinander kann interpretiert werden, sondern auch die Lage der Teilkorpora zueinander. Die relative Nähe bzw. Entfernung der Teilkorpora bietet eine Grundlage zur Abgrenzung von Phasen bzw. Clustern im Diskurs.
d) Identifikation von möglichen *zentralen Begriffen für die einzelnen Phasen/Cluster* des Diskurses: Mithilfe der hier dargelegten Verfahren kann ermittelt werden, welche Begriffe für welche Phasen/Cluster besonders charakteristisch sind. Die relative Nähe der Begriffe zu den Teilkorpora zeigt, welche Begriffe für welche Phasen/Cluster besonders charakteristisch sein können.[19]

Multivariate Analyseverfahren von Differenzbeziehungen stellen ein exploratives Tool der Diskursanalyse im Sinne des *corpus driven*-Ansatzes dar. Da Position und relative Lage der Begriffe im Koordinatensystem zueinander – im Gegensatz zu den beiden oben vorgestellten Verfahren – rein mathematisch bestimmt sind, können sie nicht direkt interpretiert werden. Position und relative Lage der Begriffe zueinander aber bieten gute Anhaltspunkte für das Finden von Begriffsrelationen in sehr großen Korpora, deren Qualität dann interpretativ durch gezieltes Nachschlagen im Korpus bestätigt werden muss.

19 Dies ist allerdings nur über einen Umweg möglich, der die Jahrgänge und Begriffe leicht gegeneinander verschiebt (Dzudzek 2008: 66).

Abbildung 8: Ergebnisse einer multivariaten Analyse von Differenzbeziehungen des Kultur-Diskurses der UNESCO



Quelle: eigene Darstellung

Die Abbildung zeigt das Ergebnis einer *multivariaten Analyse von Differenzbeziehungen* am Beispiel des Kultur-Diskurses der UNESCO (Dzudzek 2008). Grundlage der hier dargestellten Hauptkomponentenanalyse sind charakteristische Begriffe mit Raumbezug aus dem Korpus aller UNESCO-Resolutionen seit ihrer Gründung. Die Abbildung visualisiert die diskursive Verschiebung von einer Fokussierung auf den Nationalstaat in der frühen und mittleren Phase hin zur sub- und supranationalen Ebene in der jüngeren Phase, in der Begriffe wie „subregional“, „regional“, aber auch „worldwide“ und „global“ relevant werden. Auch zeigt sich anhand der räumlichen Nähe der Begriffe innerhalb der Grafik die Differenzbeziehung zwischen „colonialism“ und den „national liberation movements“, welche diskursiv mit der Befreiung der „newly independent countries“ vom Kolonialismus verknüpft werden. Dass die mathematisch bestimmte Lage der Begriffe zueinander nicht direkt interpretiert werden kann, sondern die Qualität der Beziehungen zueinander durch gezieltes Nachschlagen im Korpus überprüft werden muss, zeigt der Begriff „national culture“. Er taucht in der Grafik in der mittleren Phase auf. Dies mag zunächst verwundern, da die interpretative Analyse zeigt, dass „Nationalkultur“ das räumliche Ordnungsprinzip ist, in dem Kulturen in der ersten Phase des Diskurses verortet und verhandelt werden. Die Tatsache, dass der Begriff

aber erst überzufällig häufig in der mittleren Phase mit Begriffen wie „(neo-) colonialism“ und „national liberation movements“ auftaucht, hängt damit zusammen, dass er erst im Diskurs ausgesprochen wird, als das Konzept der nationalen Container, in denen man Kulturen verorten kann, durch die nationalen Befreiungsbewegungen der postkolonialen Staaten zum Thema gemacht wird, indem bspw. gefordert wird: „the need to reassert indigenous cultural identity and to eliminate the harmful consequences of the colonial era, […] call(s) for the preservation of national culture and traditions“ (UNESCO 2006: 60).

Exkurs: Lexikometrische bzw. korpuslinguistische Software und Ressourcen

Für lexikometrische Analysen steht eine Vielzahl von Programmen bereit, die auf den unterschiedlichen Betriebssystemen (Windows, Unix-Systeme, Linux etc.) basieren. Diese Programme wurden fast ausschließlich im Rahmen sprachwissenschaftlicher Fragestellungen entwickelt und bieten ein umfangreiches Spektrum an Funktionen. Einen Überblick über verschiedene lexikometrische Programme bieten der Linguist Noah Bubenhofer[20] sowie eine Internetdatenbank des belgischen Soziologen Christophe Lejeune[21].

Für Frequenz-, Konkordanz-, Kookkurrenzanalysen oder die Berechnung von N-Grammen stehen einerseits verschiedene Tools und Programme zur Verfügung, die ausschließlich für die jeweiligen Analyseverfahren entwickelt wurden (wie z. B. das Ngram Statistics Package (NSP) zur Berechnung von N-Grammen oder das Konkordanz-Programm AntConc). Andererseits gibt es mit dem französischsprachigen Lexico3[22] und dem englischsprachigen Wordsmith4[23] zwei komplexe lexikometrische Programme, welche die Durchführung aller oben angesprochenen lexikometrischen (Grund-)Verfahren ermöglichen. Die beiden Pogramme haben in sozialwissenschaftlichen Studien bereits mehrfach Anwendung gefunden. Für komplexe multivariate Analysen, wie bspw. Cluster- oder Faktorenanalysen, bieten sie jedoch keine bzw. nur begrenzte Möglichkeiten (Ausnahmen: SenseCluster[24], HyperBase[25]). Lexico3 bietet bspw. nur die Möglichkeit einer auf sechs Dimensionen begrenzten Faktorenanalyse.

Lexico3 und Wordsmith4 unterscheiden sich hinsichtlich ihrer Analyseverfahren nicht grundlegend. Sie ermöglichen die Berechnung von Frequenzen, Konkordanzen, N-Grammen (als *segments*

20 Online abrufbar unter http://bubenhofer.com (Zugriff: 7.10.2008)

21 Online verfügbar unter http://analyses.ishs.ulg.ac.be/logiciels/ (Zugriff: 28.1.2009)

22 Informationen zu Lexico3 finden sich unter http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/lexico3.htm (Zugriff: 18.12.2008)

23 Informationen verfügbar auf Mike Scotts Website unter http://www.lexically.net/wordsmith/index.html

24 Online verfügbar unter http://www.d.umn.edu/~tpederse/senseclusters.html (Zugriff: 14.9.2008)

25 Informationen zu dem Programm finden sich unter: http://ancilla.unice.fr/~brunet/ pub/hyperbase.html (Zugriff: 18.12.2008)

repetés bzw. *cluster* bezeichnet) und Kookkurrenzen (bei Wordsmith4 als Berechnung von keywords bezeichnet)[26]. Unterschiede gibt es jedoch bzgl. ihrer Leistungsfähigkeit und ihrer Handhabung. Wordsmith4 ermöglicht durch die direkt miteinander verknüpften drei Basis-Werkzeuge Wordlist, Concord und Keyword eine flexiblere Analyse. Zudem ist es leistungsstärker und lässt eine Analyse von Textkorpora mit einem Umfang von mehr als 150 MB zu, während Lexico3 bei diesem Datenvolumen an seine Leistungsgrenzen stößt. Vorteilhaft erscheint bei Lexico3, dass sich im Anschluss an die Analyse der N-Gramme diese mit in die Berechnung der Kookkurrenzen bestimmter Teilkorpora einbeziehen lassen. So kann Lexico3 neben den einzelnen Wörtern auch die N-Gramme berechnen, die im Korpus mit einer gewissen Signifikanz miteinander verknüpft werden. Dies ist einerseits für Analysen in Sprachen sinnvoll, in denen im Gegensatz zum Deutschen keine Bildung von zusammengeschriebenen Nominalkomposita erfolgt, wie bspw. im Fall romanischer Sprachen, in denen Komposita zumeist mithilfe einer Präposition (etwa *politique de la ville* im Französischen) oder durch Substantiv-Adjektiv-Konstruktionen (z. B. *desenvolvimento urbano* im Portugiesischen) ausgedrückt werden. Andererseits ist dies von großer Bedeutung, wenn man die regelmäßigen sprachlichen Verknüpfungen mit bestimmten, aus zwei oder mehr Lexemen bestehenden Bedeutungskonzepten untersuchen möchte, wie z. B. „Soziale Stadt“ oder „Nachhaltige Entwicklung“.

26 Je nach Analyse-Software werden unterschiedliche Signifikanztests bereitgestellt. Grundlage der Berechnung der Signifikanzen ist in beiden Fällen die absolute Häufigkeit eines bestimmten Wortes (Graphems) bzw. einer Wortfolge und die Gesamtzahl aller Wörter in einem gegebenen Korpus (Okkurrenzen). Aus dem Verhältnis zwischen der Häufigkeit dieses Wortes und der Gesamtzahl aller Wörter im Korpus lässt sich die Wahrscheinlichkeit für eine bestimmte Frequenz des Wortes in einem Teil des Korpus berechnen. Auf diese Weise lässt sich zeigen, welche Wörter und ggf. Wortfolgen in einem Teilkorpus im Vergleich zum Gesamtkorpus spezifisch häufiger bzw. seltener vorkommen. Im Fall von Lexico3 wird aus dem Verhältnis zwischen der Häufigkeit des bestimmten Wortes und der Gesamtzahl aller Wörter im Korpus die Wahrscheinlichkeit für eine bestimmte Frequenz des Wortes in einem Teil des Korpus berechnet. Dabei werden die negativen Exponenten der Zehnerpotenzen dieser Wahrscheinlichkeiten als Spezifität bezeichnet (10-x). Die empirisch gefundenen Wahrscheinlichkeiten unterschreiten dabei vielfach 10-5 deutlich. Gemäß Lebart, Salem und Berry können diese Werte daher unmittelbar als Aussage über die Spezifität eines bestimmten Wortes (Graphems) bzw. einer Wortfolge in einem bestimmten Teilkorpus interpretiert werden (1998: 135). WordSmith4 stellt zwei statistische Tests für die Berechnung der als keyness-Wert bezeichneten Signifikanzen zur Verfügung. Die Berechnung des keyness-Werts erfolgt entweder anhand des klassischen Chi-Quadrat-Tests mit Yates-Korrektur oder mithilfe des statistischen Tests auf Basis von Ted Dunnings Log Likelihood (Dunning 1993). Eine gute Einführung in diese beiden statistischen Verfahren bietet Noah Bubenhofer unter http://www.bubenhofer.com (Zugriff: 12.1.2009).

Bevor Texte mithilfe lexikometrischer Software analysiert werden können, müssen diese aufbereitet werden. So scheint es bei einigen Programmen, wie auch Lexico3, sinnvoll, den gesamten Korpus auf Kleinschreibung umzustellen, da diese Programme graphische Formen unterscheiden – dasselbe Wort einmal in Groß- und einmal in Kleinschreibung wird als zwei unterschiedliche Formen registriert. Für spezifische Fragestellungen kann es darüber hinaus sinnvoll sein, eine Lemmatisierung durchzuführen, d. h. eine Reduktion der Flexionsformen eines Wortes auf die Grundform. Für die Lemmatisierung sind sprachspezifisch spezielle Programme notwendig. WordSmith4 verfügt über eingeschränkte Möglichkeiten zur Lemmatisierung von Texten. So ist es anhand einer eigens zu erstellenden Lemma-Liste möglich, ausgewählte Flexionsformen auf die Grundform zurückzuführen.

Grundsätzlich ist festzuhalten, dass bislang keine lexikometrische Software vorliegt, die auf die Unterstützung einer sozialwissenschaftlich orientierten Diskursforschung ausgerichtet ist. So bieten die Programme Lexico3 und Wordsmith4 bspw. keine Möglichkeit, Dokumente zu verwalten und auf diese Weise flexibel unterschiedliche Kombinationen von Dokumenten (z. B. Presseartikel unterschiedlicher Jahrgänge oder mit unterschiedlichen Schlagwörtern) kontrastieren zu können. Folglich muss für neue Kontrastierungen jeweils aufwendig ein neuer Korpus erstellt werden. Hilfreich für eine flexible Korpuserstellung erweisen sich Datenbankensysteme, mit denen sich die einzelnen digitalen Texte verwalten lassen.[27]

Das Institut für Deutsche Sprache (IDS) sowie die Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) bieten digitale Korpora, die über das Internet abfragbar sind. Verschiedene lexikometrische Analyseverfahren können mit diesen Ressourcen online durchgeführt werden wie bspw. Kookkurrenzanalysen und die Berechnung von Kookkurrenzprofilen oder Synonymen (Cosmas II[28], CCBD[29], DWDS[30]) – ohne dass dafür spezielle Programme notwendig wären.

27 Aus Sicht einer sozialwissenschaftlichen Diskursforschung erscheint es daher wünschenswert und wichtig, dass die komfortableren QDA-Programme (bspw. MaxQDA und Atlas.ti) zukünftig um korpuslinguistische Werkzeuge ergänzt werden.

28 Über Cosmas II kann anhand eines umfangreichen Abfragepakets auf die Korpora des IDS zugriffen werden, verfügbar unter http://www.ids-mannheim.de/cosmas2/uebersicht.html (Zugriff: 18.12.2008).

29 Zugriff auf die Kookkurrenzdatenbank CCBD des IDS unter http://corpora.ids-mannheim.de/ccdb (Zugriff: 18.12.2008)

30 Das Digitale Wörterbuch der Deutschen Sprache (DWDS) der Berlin-Brandenburgischen Akademie der Wissenschaften bietet eine lemmabasierte Kollokationssuche im DWDS-Kerncorpus, durchführbar unter http://www.dwds.de (Zugriff: 18.12.2008).

Fazit: Potenziale und Grenzen korpuslinguistisch-lexikometrischer Verfahren für die Diskursforschung

Wie die Ausführungen und Beispiele gezeigt haben, können lexikometrische Analysen im Rahmen diskursanalytischer Arbeiten einen wichtigen Beitrag leisten. Sie ermöglichen es, große Textmengen zu erfassen und auf Regelmäßigkeiten und Strukturen zu untersuchen, die „von Hand", d. h. durch Lesen des/der Forschenden, nicht zu erfassen wären. Lexikometrische Verfahren bieten zudem die Chance, induktiv diskursive Strukturen herauszuarbeiten, die gerade nicht den Vorannahmen der Forschenden entsprechen.

Strukturalistische bzw. poststrukturalistische Theorien teilen die Auffassung, dass Bedeutung durch regelmäßige Verknüpfungen von symbolischen (insbesondere sprachlichen) Formen entsteht. Die Lexikometrie hilft, diese theoretische Annahme zu operationalisieren, indem sie Differenzbeziehungen von sprachlichen Elementen untersucht und damit die kontextspezifische Konstitution von Sinn im diachronen oder synchronen Vergleich herausarbeitet. Gerade der Vergleich unterschiedlicher Teilkorpora kann dabei eingesetzt werden, um auch Unterschiede, Verschiebungen und Brüche innerhalb des Diskurses – etwa Veränderungen über die Zeit oder Unterschiede zwischen Sprecherpositionen – herauszuarbeiten. Damit kann aufgezeigt werden, wie sich die Konstitution von Bedeutungen abhängig vom jeweiligen diskursiven Kontext verschiebt, wie sie verändert und von neuen Formen der Sinnproduktion herausgefordert wird (Glasze 2007a, b). Mithilfe der entsprechenden statistischen Verfahren können aber nicht nur Unterschiede zwischen einzelnen Teilkorpora untersucht werden, sondern auch Begriffshäufungen im Kontext bestimmter sprachlicher Formen. So kann für Fragestellungen, die sich mit der Herstellung kollektiver Identität und diskursiver Gemeinschaften beschäftigen, nach Kookkurrenzen des Begriffs „wir" (sowie „uns" etc.) gesucht werden. Die Signifikanten, die in solchen sprachlichen Kontexten besonders häufig auftreten, können Hinweise auf Prozesse der Identifikation und Abgrenzung bieten (Mattissek 2007).

Die Verwendung lexikometrischer Verfahren stößt allerdings auch an Grenzen: So kann mittels lexikometrischer Verfahren bspw. gezeigt werden, ob und wann das Wort „Afrika" regelmäßig mit „Armut" verknüpft und dementsprechend eine bestimmte Bedeutung hergestellt wird – oder nicht. Die lexikometrischen Analysen sind aber nur teilweise in der Lage, die Qualität dieser Verknüpfungen herauszuarbeiten und damit zu analysieren, ob bspw. zwischen den Elementen Beziehungen der Temporalität, der Äquivalenz, der Opposition oder der Kausalität herge-

stellt werden. Es erscheint daher heuristisch fruchtbar, auch diese Dimension der Konstitution von Bedeutung ins Blickfeld zu nehmen.[31] Darüber hinaus erweist sich die Lexikometrie auch als wenig hilfreich, wenn es darum geht, ungesagtes oder implizites Wissen (etwa Prämissen, die als selbstverständlich vorausgesetzt werden) zu erfassen. Ebenso wenig lassen sich mit ihrer Hilfe Phänomene wie Ironie oder Sarkasmus analysieren.

In der Regel bietet es sich daher in empirischen Arbeiten an, lexikometrische Methoden mit anderen Verfahren zu kombinieren, die die Konstitution von Bedeutung in einzelnen Aussagen oder Texten adressieren. Insbesondere können die hier vorgestellten Verfahren der quantitativen Makroanalyse von Texten sinnvoll mit Verfahren der Aussagen- und Argumentationsanalyse und kodierenden Verfahren verknüpft werden (Kap. 12, 13 und 14). Für diese „Mikroverfahren" liefert die Lexikometrie wichtige Anregungen, indem sie Hinweise auf relevante Themen- und Begriffsfelder gibt.

Exkurs: Lexikometrische bzw. korpuslinguistische Fachbegriffe

Signifikanz/Signifikanztests: In der Lexikometrie bzw. Korpuslinguistik spielen Signifikanztests eine wichtige Rolle. Sie werden verwendet, um bspw. zu überprüfen, ob ein bestimmtes Lexem in einem Teilkorpus signifikant häufiger vorkommt als in einem Referenzkorpus. In diachroner Perspektive kann damit geprüft werden, ob die Verwendung eines Sprachmusters sich zeitlich signifikant verändert. Auch bei der Berechnung von Kookkurrenzen und N-Grammen werden Signifikanztests eingesetzt, um zu bestimmen, ob zwei Wörter überzufällig (signifikant) häufig zusammen auftreten. Regelmäßig verwendete Signifikanztests sind bspw. der Chi-Quadrat-Test oder der Log-Likelihood-Test.

Multivariate Verfahren: Ziel multivariater Analyseverfahren von Differenzbeziehungen ist die Erweiterung der Kookkurrenzanalyse hin zu einer Analyse eines komplexen differenziellen Netzes von Zeichen (Textelementen), in dem Bedeutung konstituiert wird. Dabei wird der Begriff Kookkurrenz um eine Vielzahl von Dimensionen erweitert, so dass er die Beziehung *mehrerer* Elemente innerhalb eines differenziellen Zeichensystems beschreibt. Mithilfe multivariater Dimensionsreduktionsverfahren lässt sich dieser komplexe n-dimensionale Zusammenhang in einem zwei-dimensionalen Koordinatensystem visualisieren.

31 Prinzipiell könnten diese Fragen auch mittels lexikometrischer Verfahren adressiert werden – die Untersuchung müsste dann letztlich unendlich lange fortgesetzt werden, um auch die signifikanten Umgebungen der signifikanten Umgebungen der signifikanten Umgebungen zu untersuchen usw. In der Forschungspraxis ist dies allerdings kaum umsetzbar.

corpus based: Korpuslinguistische Verfahren, bei denen die Verteilung eines im Voraus definierten lexikalischen Elements in einem Textkorpus untersucht wird.

corpus driven: Induktive korpuslinguistische Verfahren, die ohne im Voraus definierte Suchanfragen auskommen und damit die Chance bieten, auf Strukturen zu stoßen, an die man nicht schon vor der Untersuchung gedacht hat.

Genre: Mit dem Begriff des Genres werden in den Sprachwissenschaften Gruppen von Texten bezeichnet, für deren Strukturierung und damit deren Kohärenz sich historisch spezifische, institutionell stabilisierte Regeln etabliert haben. So gelten für die Strukturierung und Kohärenz wissenschaftlicher Fachaufsätze andere Regeln als für Zeitungsartikel und wiederum andere für politische Reden.

Graphem: Als Graphem werden die kleinsten bedeutungsunterscheidenden Einheit der geschriebenen Sprache bezeichnet.

Kollokationen: s. unter Kookkurrenzen

Kookkurrenzen: Als Kookkurrenz wird das überzufällig häufige gemeinsame Auftreten zweier oder mehrerer Wörter in einer bestimmten definierten Umgebung eines bestimmten Schlüsselwortes bezeichnet. Die Untersuchung von Kookkurrenzen zeigt, welche Wörter und Wortfolgen im Korpus mit einer gewissen Signifikanz miteinander verknüpft werden, d. h. das gemeinsame Auftreten ist höher, als bei einer Zufallsverteilung aller Wörter erwartbar wäre. Eine Kookkurrenzanalyse untersucht die Wortumgebung eines ausgewählten Begriffs und gibt dadurch Aufschluss über seine Bedeutungskonstitution.

Lexem: Das Lexem ist die kleinste semantische Einheit. Ein Lexem bezeichnet eine Menge von Wörtern, welche alle Flexionsformen des gleichen Grundwortes darstellen, d. h. sich nur in bestimmten morphosyntaktischen Merkmalen (Kasus, Numerus, Tempus usw.) unterscheiden. So gehören z. B. die verschiedenen Flexionsformen eines Substantivs oder Verbs zum selben Lexem (laufen, läuft, läufst = ein Lexem; laufen, Läufer = zwei Lexeme).

Lemma/Lemmatisierung: Ein Lemma bezeichnet in der Lexikographie und Linguistik eine lexikographische Standard-Notation für ein Lexem. Ein Lemma ist die Grundform eines bestimmten Wortes, die nach bestimmten Notationskonventionen gebildet wird (z. B. im Deutschen für Nomen die Verwendung des Nominativ Singular, für das Verb der Infinitiv). Anhand einer Lemmatisierung wird eine Reduktion der Flexionsformen eines Wortes auf die Grundform durchgeführt.

Lexikometrie: Die Lexikometrie zielt darauf ab, großflächige Strukturen der Sinn- und Bedeutungskonstitution in Texten zu erfassen. Lexikometrische Verfahren untersuchen die quantitativen Beziehungen zwischen lexikalischen Elementen in geschlossenen Textkorpora, d. h. in Textkorpora, deren Definition, Zusammenstellung und Abgrenzung klar bestimmt ist und die im Laufe der Untersuchung unverändert bleiben.

Okkurrenzen: In der Linguistik bezeichnet man mit Okkurrenz die Häufigkeit, mit der ein bestimmtes sprachliches Element wie bspw. ein Wort in einem komplexeren sprachlichen Zusammenhang auftritt.

Konkordanz: Eine Konkordanz ist eine Liste, die alle Vorkommen eines ausgewählten Wortes – oder auch Wortfolgen – in seinem Kontext zeigt. Für Konkordanzen üblich ist eine zeilenweise Darstellung, die als KWIC *(key word in context)* bezeichnet wird. Auf dessen linker und rechter Seite wird ein festgelegter Kontext, bestehend aus einer bestimmten Anzahl an Zeichen oder Wörtern, angezeigt. Konkordanzanalysen können sinnvoll als Vorbereitung und Hilfe für die qualitative Interpretation des Kontextes bestimmter Schlüsselwörter verwendet werden.

Sprecherposition: Sprecherpositionen werden in Anlehnung an Überlegungen Foucaults als institutionell stabilisierte Positionen i. d. R. innerhalb von Organisationen gefasst, die spezifische Zugangskriterien haben und die bestimmte Möglichkeiten, Tabus und Erwartungen des Sprechens bzw. allgemein der Textproduktion mit sich bringen – weitgehend unabhängig von den Individuen, welche die Position einnehmen. Die Sprecherpositionen sind dabei selbst diskursiv konstituiert.

Textkorpus: Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen, die als empirische Grundlage für sprachwissenschaftliche Untersuchungen dient. Die Beschaffenheit des Korpus hängt von der spezifischen Fragestellung und der methodischen Herangehensweise der Untersuchung ab. Die typischerweise digitalisierten Daten können ggf. Metadaten enthalten, die diese Daten beschreiben, sowie linguistische Annotationen.

N-Gramme: Ein N-Gramm ist eine Folge aus N Zeichen oder N Wörtern. So besteht bspw. ein Monogramm aus *einem* Zeichen, bspw. nur aus einem einzelnen Buchstaben, das Bigramm aus *zwei* und das Trigramm aus *drei* Zeichen. Im Rahmen von lexikometrischen Analysen bezieht sich die Berechnung von N-Grammen vor allem auf Wortkombinationen aus N Wörtern (auch *Multi Word Units* oder *segments repetés* genannt). So lassen sich N-Gramme nach ihrer Frequenz sortieren oder in komplexeren Verfahren nach ihrer Signifikanz gewichten.

Literatur

Atteslander, Peter/Cromm, Jürgen (2006): Methoden der empirischen Sozialforschung, Berlin: Schmidt.

Baker, Paul (2006): Using corpora in discourse analysis, London/New York: Continuum.

Baker, Paul/Gabrielatos, Costas/Khosravinik, Majid/Kryzanowski, Michal/Mcenery, Tony/Wodak, Ruth (2008): A useful methodological synergy? Combining critical discourse analysis and corpus lin-

guistics to examine discourses of refugees and asylum seekers in the UK press. Discourse & Society 19 (3), S. 273–306.

Belica, Cyril/Steyer, Kathrin (2005): Korpusanalytische Zugänge zu sprachlichem Usus, Prag: Karolinum Verlag.

Berelson, Bernard (1952): Content analysis in communication research, New York: Hafner Press.

Bonnafous, Simone (2002): Analyse de contenu. In: Charaudeau, Patrick/Maingueneau, Dominique (Hg.), Dictionnaire d'analyse du discours, Paris: Éd. du Seuil, S. 39–41.

Bonnafous, Simone/Tournier, Maurice (1995): Analyse du discours, lexicométrie, communication et politique. Langages 117, S. 67–81.

Brailich, Adam/Germes, Mélina/Glasze, Georg/Pütz, Robert/Schirmel, Henning (2009): Die diskursive Konstitution von Großwohnsiedlungen in Frankreich, Deutschland und Polen. Europa Regional 17 (im Druck).

Bubenhofer, Noah (2008): „Es liegt in der Natur der Sache ...". Korpuslinguistische Untersuchungen zu Kollokationen in Argumentationsfiguren. In: Mellado Blanco, Carmen (Hg.), Beiträge zur Phraseologie aus textueller Sicht, Hamburg: Kovac, S. 53–72.

Busse, Dietrich/Teubert, Wolfgang (1994): Ist der Diskurs ein sprachwissenschaftliches Objekt? Zur Methodenfrage der historischen Semantik. In: Busse, Dietrich/Hermanns, Fritz/Teubert, Wolfgang (Hg.), Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik, Opladen: Westdeutscher Verlag, S. 10–28.

Dunning, Ted (1993): Accurate Methods for the Statistics of Surprise and Coincidence. Computial Linguistics 19 (1), S. 61–74.

Dzudzek, Iris (2008): Kulturelle Vielfalt versus kulturelle Hegemonie. Eine diskursanalytische Untersuchung kultur-räumlicher Repräsentationen und identitätspolitischer Kämpfe in der UNESCO (= unveröff. Diplomarbeit am Geographischen Institut der Universität Münster), Münster.

Fiala, Pierre (1994): L'interprétation en lexicométrie. Une approche quantitative des données lexicales. Langue française (103), S. 113–122.

Foucault, Michel (1973 [1969]): Archäologie des Wissens, Frankfurt a. M.: Suhrkamp

Glasze, Georg (2007a): The discursive constitution of a world spanning region and the role of empty signifiers: the case of Francophonia. Geopolitics 12 (4), S. 656–679.

Glasze, Georg (2007b): Vorschläge zur Operationalisierung der Diskurstheorie von Laclau und Mouffe in einer Triangulation von lexiko-

metrischen und interpretativen Methoden. FQS – Forum Qualtitative Sozialforschung 8 (2). Online unter http://www.qualitative-research.net/index.php/fqs/article/view/239, abgerufen am 1.2.2009.

Glasze, Georg/Germes, Mélina/Weber, Florian (2009): Krise der Vorstädte oder Krise der Gesellschaft? Geographie und Schule (177), S. 17–25.

Guilhaumou, Jacques (1997): L'analyse de discours et la lexicometrie. Le pere duchesne et le mouvement cordelier (1793–1794), Lexicometrica. Online unter http://www.cavi.univ-paris3.fr/lexicometrica/article/numero0/jgadlex.htm.

Guilhaumou, Jacques (2003): Geschichte und Sprachwissenschaft – Wege und Stationen (in) der 'analyse du discours'. In: Keller, Reiner/Hirseland, Andreas/Schneider, Werner/Viehöver, Willy (Hg.), Handbuch Sozialwissenschaftliche Diskursanalyse, Opladen: Leske + Budrich, S. 19–65.

Jung, Matthias (1994): Zählen oder deuten? Das Methodenproblem der Diskursgeschichte am Beispiel der Atomenergiedebatte. In: Busse, Dietrich/Hermanns, Fritz/Teubert, Wolfgang (Hg.), Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik, Opladen: Westdeutscher Verlag, S. 60–81.

Koteyko, Nelya (2006): Corpus linguistics and the study of meaning in discourse. The Linguistics Journal 1 (2), S. 132–157.

Laclau, Ernesto (1990): New reflections on the revolution of our time, London: Verso.

Laclau, Ernesto/Mouffe, Chantal (1985): Hegemony & socialist strategy: towards a radical democratic politics, London: Verso.

Lebart, Ludovic/Salem, André (1994): Statistique textuelle, Paris: Dunod.

Lebart, Ludovic/Salem, André/Berry, Lisette (1998): Exploring textual data, Dordrecht: Kluwer.

Lemnitzer, Lothar/Zinsmeister, Heike (2006): Korpuslinguistik: eine Einführung, Tübingen: Narr.

Lüsebrink, Hans-Jürgen (1998): Begriffsgeschichte, Diskursanalyse und Narrativität. In: Reichardt, Rolf (Hg.), Aufklärung und Historische Semantik. Interdisziplinäre Beiträge zur westeuropäischen Kulturgeschichte, Berlin: Duncker & Humblot, S. 29–44.

Maingueneau, Dominique (1991): L'analyse du discours: introduction aux lectures de l'archive, Paris: Hachette.

Maingueneau, Dominique (2000 [1986]): Linguistische Grundbegriffe zur Analyse literarischer Texte, Tübingen: Narr.

Marchand, Pascal (1998): L'analyse du discours assistée par ordinateur: Concepts, méthodes, outils, Paris: Colin.

Marchart, Oliver (1998): Einleitung: Undarstellbarkeit und „ontologische Differenz". In: Marchart, Oliver (Hg.), Das Undarstellbare der Politik. Zur Hegemonietheorie Ernesto Laclaus, Wien: Turia + Kant, S. 7–22.

Mattissek, Annika (2007): Diskursive Konstitution städtischer Identität – Das Beispiel Frankfurt am Main. In: Berndt, Christian/Pütz, Robert (Hg.), Kulturelle Geographien. Zur Beschäftigung mit Raum und Ort nach dem Cultural Turn, Bielefeld: transcript, S. 83–111.

Mattissek, Annika (2008): Die neoliberale Stadt. Diskursive Repräsentationen im Stadtmarketing deutscher Großstädte, Bielefeld: transcript.

Mayaffre, Damon (2004): Formation(s) discursive(s) et discours politiques: l'exemplarité des discours communistes versus bourgeois durant l'entre-deux-guerres. Texto ! (06/2004). Online unter http://www.revue-texto.net/index.php?id=585.

Mayring, Philipp (2008 [1983]): Qualitative Inhaltsanalyse. Grundlagen und Techniken, Weinheim/Basel: Beltz.

Merten, Klaus (1995): Inhaltsanalyse: Einführung in Theorie, Methode und Praxis, Opladen: Westdeutscher Verlag.

Niehr, Thomas (1999): Halbautomatische Erforschung des öffentlichen Sprachgebrauchs oder Vom Nutzen computerlesbarer Textkorpora. Zeitschrift für germanistische Linguistik 27, S. 205–214.

Orpin, Debbie (2005): Corpus linguistics and critical discourse analysis: Examining the ideology of sleaze. International Journal of Corpus Linguistics 10 (1), S. 37–62.

Reichardt, Rolf (1998): Historische Semantik zwischen lexicométrie und new cultural history. In: Reichardt, Rolf (Hg.), Aufklärung und Historische Semantik. Interdisziplinäre Beiträge zur westeuropäischen Kulturgeschichte, Berlin: Duncker & Humblot, S. 7–28.

Saussure, Ferdinand de (1931 [1916]): Grundfragen der allgemeinen Sprachwissenschaft, Berlin: de Gruyter.

Scherer, Carmen (2006): Korpuslinguistik, Heidelberg: Winter.

Ter Braak, Cajo J. F./Smilauer, Petr (2002): CANOCO Reference manual and CanoDraw for Windows User's guide. Software for Canonical Community Ordination (version 4.5). Microcomputer Power: Ithaca (USA).

Teubert, Wolfgang (1999): Korpuslinguistik und Lexikographie. Deutsche Sprache (4), S. 293–313.

Teubert, Wolfgang (2005): My version of corpus linguistics. Journal of Corpus Linguistics (1/2005), S. 1–13.

Tognini-Bonelli, Elena (2001): Corpus linguistics at work, Amsterdam: Benjamins.

UNESCO (2006): Resolution 18C – Paris 1974. In: UNESCO (Hg.), Resolutions and Decisions 1946–2005 (CD-ROM), Paris.

Wernet, Andreas (2006): Einführung in die Interpretationstechnik der Objektiven Hermeneutik, Wiesbaden: VS Verlag für Sozialwissenschaften.

Williams, Glyn (1999): French discourse analysis. The method of post-structuralism, London/New York: Routledge.